面向网络大数据的信息检索与挖掘

郭嘉丰 程学旗

摘要：随着网络数据的爆炸性增长，信息处理技术面临着前所未有的巨大挑战。如何从体量巨大、增长迅速、结构复杂、良莠不齐的数据中发掘潜在价值成为了关键难题。面向网络大数据的信息检索与挖掘技术，旨在通过对大数据的深度分析与建模，有效弥合用户需求与网络数据之间的信息鸿沟。本文介绍了面向网络大数据的深度检索与挖掘的一系列关键技术，包括用户查询理解与处理、文档建模与理解及检索模型等。

关键词：信息检索 数据挖掘 查询理解 话题模型 排序学习

1 引言

伴随着互联网技术的迅猛发展和普及，用户规模的爆发式增长，互联网已经步入了“大数据”时代。网络大数据的“大”，不仅仅体现在其体量巨大（大数据的起始计量单位至少是 Petabyte[1]、Exabyte[2]或 Zettabyte[3]），而且还体现在其增长异常迅猛（通常是指数级的速率），数据类型多样（包括了文本、图像、声音、视频等等），数据质量良莠不齐并且关联关系复杂。同时，网络大数据另外一个突出的特点就是其价值密度低，大数据中包含了大量重复、噪声和垃圾数据，存在大量共现但又毫无意义的关联模式，如果缺乏有效的信息处理手段提取网络大数据中潜在的价值，那么网络大数据不仅不能成为一个价值“宝藏”，反倒会成为一个数据的“坟墓”。

信息检索和挖掘是网络信息处理的关键技术。网络大数据对于信息检索和挖掘技术而言是一把双刃剑：一方面，网络大数据提供了需要检索和挖掘技术来处理的丰富的数据源，大规模的样本资源可以更好地支持文本分析、关系挖掘、用户语义理解、图像处理等等关键技术的发展；但另一方面，网络大数据复杂的内在特征对传统搜索与挖掘技术提出了严峻的挑战。例如网络大数据越来多地存在于电商、问答等私有化网络或者深网中，包括了结构化数据、半结构化数据和非结构化数据，使得数据的获取和存储更加困难；数据庞大的规模、复杂的关联关系，使得传统的分析和挖掘技术在计算的时空复杂度上激增；另外迅猛的数据增长速率，巨大的数据体量也使得传统的全量计算模式（依赖于全体样本的计算模式）不再适用。为了能够有效地克服这些难题，充分挖掘和利用网络大数据中的信息价值，我们需要研究面向网络大数据的深度检索与挖掘技术。

面向网络大数据的深度检索与挖掘技术，旨在通过对大数据的深层分析与建模，有效弥合用户需求与网络数据之间的鸿沟，帮助用户有效发现和准确定位所需的信息。面向网络大数据的深度检索与挖掘技术主要需要解决好三个方面的问题，即在用户空间对用户查询的深度理解与处理，在数据空间对文档等数据的深度建模与理解，以及对用户查询和网络数据的高效智能匹配。

首先，用户查询是用户对自己信息需求的主要表达手段。要能够在网络大数据环境下实现有效的信息定位，需要能够对用户查询进行更深层次的理解与处理，以便更加准确地捕获

[1] 10^{15} 千万亿
[2] 10^{18} 百亿亿
[3] 10^{21} 十万亿亿

用户的意图，从语义层面更好地实现信息的匹配。搜索引擎十多年的发展积累了大量的用户查询日志，而且这样的日志数据还在持续的增长（例如据统计百度每天处理的搜索查询量超过了 50 亿次）。海量的用户查询日志数据提供了异构、丰富的用户行为数据，为实现查询的深层理解与处理提供了坚实基础。解决如何基于海量用户查询日志数据，在更深层次解析查询结构，度量查询的相关关系，分析查询效用等一系列问题，成为了理解用户的查询意图及对其建模的关键。

其次，网络数据对象（如文档、图片等）是信息检索和挖掘的对象，对其合理的建模和分析，提取其中的关键语义、关键模式，才能有效地发掘数据内含的规律和潜在价值，实现更准确的信息获取。特别是在大规模网络文本数据环境，提取其中的语义话题是数据挖掘和检索的一项关键技术。然而，传统的话题建模的基本假设（所有文档都共享同样的话题维度）不再适用于网络大数据。实际网络大数据具有特征稀疏、语义稀疏等特点，大规模文本数据集潜在包含了高维话题但同时单个文本却只有极少话题。这些问题都对网络大数据的话题建模技术提出了新的挑战和需求。

最后，检索模型旨在解决用户需求空间和网络数据空间的智能匹配。当前排序学习技术由于其坚实的理论基础、灵活的建模方式和优异的排序性能，成为了学术界和工业界主流的检索模型。然而，传统的排序学习技术依赖于对全集样本的多级标注和学习，标注代价高且不能很好地体现检索中关注位置的特点。如何提高排序学习技术在大数据下的性能，构建适用于网络大数据的排序学习标注、建模和评价体系，成为了一个非常实际的课题。

本文将从用户查询意图的理解，网络文档稀疏话题建模，以及大数据下的排序学习三个方面介绍我们在网络大数据信息检索和挖掘技术方面近年来取得的重要研究成果。其中，第二节主要就从基于大规模用户查询日志的查询结构分析、查询相关关系度量、查询效用分析等方面介绍查询理解方面的相关成果；第三节从语义稀疏和特征稀疏两个方面介绍大数据稀疏话题建模方法；第四节介绍大数据下的高性能 Top-k[4]排序学习技术；最后，在第五节进行总结。

2 基于大规模用户查询日志的查询理解

用户查询理解旨在通过对用户查询的建模、分析和处理，理解用户查询的意图，提高信息检索的质量和用户体验。大规模用户查询日志为深层次理解用户查询提供了基础且宝贵的数据。本节基于用户查询日志从查询串本身、查询之间以及查询会话序列三个层次展开研究，提出了查询结构解析[1]、查询相似关系度量[2]、基于效用的查询推荐[3]等模型与方法，逐层深入地理解用户查询的意图并进行处理。

2.1 基于命名实体识别的用户查询结构解析

为了能够理解用户查询的语义和意图，我们对查询结构进行了分析和建模。通过研究发现，大约有 71% 的用户查询包含命名实体，而这些命名实体通常代表了用户检索的核心语义。识别这些查询中的命名实体将可以帮助我们更好地理解用户检索的意图，从而更好地辅助检索。例如，在相关检索中，我们可以通过对查询中的实体和其他部分分别赋予不同的权重来提高排序的质量；在查询推荐时，查询中命名实体的类别信息则可以帮助我们产生更加相关和多样的查询建议。例如，对于“harry potter walkthrough”这个查询，我们可以通过分析其中的实体和上下文信息，发现该查询属于对游戏类别的查询，同时我们可以利用游戏类

[4] 在某个数据集中找出按某种方式排序的前 k 名成员的算法

别的其它查询上下文来产生推荐，如“harry potter cheats”。正因如此，我们首次提出了查询中命名实体识别这一研究问题。

在这个研究工作中，我们提出了利用大规模用户查询日志和一个新颖的概率框架来进行查询结构分析，识别用户查询中的命名实体及其上下文模板。不失一般性，包含单个命名实体的查询可以表示为一个三元组 (e，t，c)，其中 e 代表命名实体，t 代表查询上下文，c 代表 e 的类别。查询中命名实体识别的问题就转化为给定查询 q，我们需要为其寻找具有最大联合概率 Pr(e, t, c) 的最优三元组 (e, t, c)，即

$$\begin{aligned}(e,t,c)^* &= \text{argmax}_{(e,t,c)} \Pr(q,e,t,c) \\ &= \text{argmax}_{(e,t,c)} \Pr(q|e,t,c)\Pr(e,t,c) \\ &= \text{argmax}_{(e,t,c)\in G(q)} \Pr(e,t,c)\end{aligned}$$

我们发现，这个联合概率可以进一步分解并利用大规模用户查询日志和一个话题模型来进行估计。在这里，使用话题模型的一个特殊挑战在于查询的语义类别（对应话题模型的隐藏话题）是预先定义的，而传统的无监督话题模型学习得到的隐藏话题无法准确地和预定义的查询语义类别进行对齐。因此，我们提出了一个基于潜在狄利克里分布的弱指导学习的话题模型，称为 WS-LDA（Weakly Supervised Latent Dirichlet Allocation），并把它应用于我们的实体识别问题中。WS-LDA 不同于使用无指导学习的传统 LDA[4]，其目标函数如下式所示：

$$\begin{aligned}O(D,Y|\Theta) &= \sum_{d=1}^{M} O(w_d, y_d|\Theta) \\ &= \sum_{d=1}^{M} \log\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)\, p(w_{dn}|z_{dn},\beta)\right)d\theta_d)) + \sum_{d=1}^{M}\lambda\sum_{i=1}^{K} y_{di}\bar{z}_{di}\end{aligned}$$

其中前半部分是传统 LDA 模型下的似然函数，后半部分是弱约束条件以实现话题对齐。WS-LDA 可以利用人工标注的弱指导信息来指导话题模型学习。

我们在随机采样的 1200 万条用户查询的数据集上进行实验，结果表明，我们提出的方法在前 3 个识别结果上可以达到 97.5%的准确率。基于该方法的识别结果，我们可以提高信息检索的排序性能，排序性能相比于原始相关排序结果在 NDCG@3[5]指标上可以提高 4.4%。

2.2 基于意图感知的用户查询相似度度量

对高维稀疏的用户查询进行相似度度量是用户查询意图理解的一个核心问题，它可以广泛地用于查询分类聚类、查询扩展、查询推荐等相关应用。然而由于用户查询往往非常简短，而且语义模糊、意图多样，对于查询相似度的度量并非是一件简单的事情。例如，给定用户查询“Apple”，假如用户的查询意图是查找水果，那么它就和“Apple tree”这样的查询相似；而假如用户的查询意图是查找苹果公司的产品，则它将和“Apple store”这样的查询相似。由于查询意图不同，这样的相似度是不能够相互比较的。也就是说，我们并不能够得出如“Apple tree”比“Apple store”更加相似于“Apple”，或者反之“Apple store”比“Apple tree”更加相似于“Apple”的结论；更不能得出由于“Apple tree”和“Apple”相似，“Apple store”也和“Apple”相似，所以“Apple tree”和“Apple store”相似这样的结论。然而传

[5] Normalized Discounted Cumulative Gain，归一化折扣累积增益，一种信息检索研究领域广泛应用的评估测度

统的查询相似度的度量基于查询单一的特征表达，采用单一的度量尺度进行计算，它们或者是基于对的方法（pair-wise method），或者是基于图的方法（graph-based method），这些方法就会产生上述例子中所述的错误或者不恰当的相似关系度量结果。

在本文中，我们首次提出了意图感知的查询相似度度量这个概念，即查询相似度需要定义在查询意图之上，只有这样我们才能够得到更加准确的相似度度量，避免产生传统度量准则的种种问题。在查询意图感知的查询相似度度量方法中，我们利用查询的搜索结果以及大规模用户查询日志中的点击日志数据，来学习查询的潜在意图。我们使用了一个正则化话题混合模型来为用户查询的潜在意图建模。该模型可以充分利用上述两类数据来进行意图的学习与推断，它目标是最大化一个正则化的似然函数，其中 λ 是一个正则化因子：

$$\begin{aligned} L &= L - \lambda R \\ &= \sum_{i=1}^{N}\sum_{j=1}^{M} n(q_i, w_j)\log(P(q_i)\sum_{k=1}^{K} P(w_j|s_k)P(s_k|q_i)) - \lambda \sum_{i,j=1}^{N}\sum_{k=1}^{K} C_{ij}(P(s_k|q_i) - Ps_k|q_j))^2 \end{aligned}$$

在我们学得的模型基础之上，可以对每个查询抽取其基于查询意图的不同的表达形式，即意图感知的表达。基于意图感知的表达，我们可以采用传统的配对方法[5]以及基于图的方法[6]来进行意图感知的查询相似度计算。我们提出的意图感知的查询相似度度量，不仅可以更加精确地计算查询之间的相似度，也可以很好地适应如结构化的查询推荐、检索结果的多样化等实际应用。

我们在大规模用户查询日志数据集上进行实验，验证我们提出的查询意图感知的相似度的有效性。我们通过人工标注获得了 200 个具有歧义的查询以及其对应于不同意图下的代表性相似查询，通过大规模点击日志进行意图学习和度量。一个好的查询相似度度量，可以让相同意图的查询相似度尽可能大，让不同意图的查询相似度尽可能小。因此，我们利用类间相似度和类内相似度的期望比例(H Score)作为评价指标。实验结果表明，我们提出的方法，在相似度度量评价指标上可以显著地优于传统的配对方法和基于图的方法（见表 1）。

表1. 不同相似度度量方法下的类间-类内相似度比例

方法	类内-类间相似度比例Hs(Sim)	显著性差异
单词余弦相似度 [1]	0.47±0.06	> Embed-Click***
意图感知余弦相似度 [2]	0.08±0.03	> Cos-Word ***> Embed-Click***
点击嵌入法 [3]	0.54±0.02	
意图感知嵌入法 [4]	0.09±0.03	> Cos-Word ***> Embed-Click***

[1] Cos-Word； [2] Cos-Intent； [3]Embed-Click； [4] Embed-Intent

***表示显著性水平为 0.01；

2.3 基于效用分析的用户查询推荐

为了有效地帮助人们表达查询意图，查询推荐成为了搜索引擎核心工具。然而现有的查询推荐方法主要在查询词这个层面上向用户推荐相关性查询或差异性查询[7,8]，并没有在查询结果层面上来考虑推荐的真正目的，即通过推荐帮助用户找到期望的信息。为解决这一问题，我们在本文的工作中首次提出了向用户推荐高效用性的（Utility）查询，使查询能够更好地满足用户的信息需求。查询的效用性定义为：用户能够从该查询的检索结果中获得的有用信息量。值得注意的是，尽管在一些查询推荐的研究工作中提到了效用性的概念，但这些工作与我们的研究工作有本质的区别。

效用性查询推荐研究最大的挑战在于：如何挖掘各个查询的效用。我们通过分析发现，用户的搜索行为，尤其是用户的查询重构行为和查询点击行为，包含了大量有价值的查询效用信息。进一步分析发现，查询的效用包含两个部分，即感知效用（Perceived Utility）和后验效用（Posterior Utility）:（1）感知效用是指用户是否对该查询的搜索结果感兴趣，只有感兴趣用户才会进一步点击这些结果并查看其内容;（2）后验效用是指所点击的结果是否能够满足用户的信息需求。查询的效用最终定义为这两个组成效用的乘积。

我们提出了一个基于动态贝叶斯网络的查询效用模型(Query Utility Model，QUM）来进行查询效用的学习（图 1)。其中，R_i 表示在位置 i 是否存在查询重构，C_i 表示用户是否点击第 i 个位置重构查询的搜索结果，A_i 表示第 i 个位置重构查询的搜索结果是否吸引用户，S_i 表示在位置 i，用户的信息需求是否得到满足。在我们的模型中，与查询效用相关的参数有两个，分别是感知效用 α 和后验效用 β，这两个量可以通过极大似然估计获得。

实验在公开的 UFindit 数据集上进行，基于两个自动评测的指标：查询相关率（Query Relevant Ratio，QRR））和相关文档均值（Mean Relevant Document，MRD））进行评价。为了验证我们提出的查询效用性模型方法的效果，我们将其与当前流行的基于会话和查询流程图的推荐方法相比。评测结果显示，我们提出的查询效用性模型方法的推荐结果在查询相关率和意义相关文档两个评测指标下均有最好的结果。

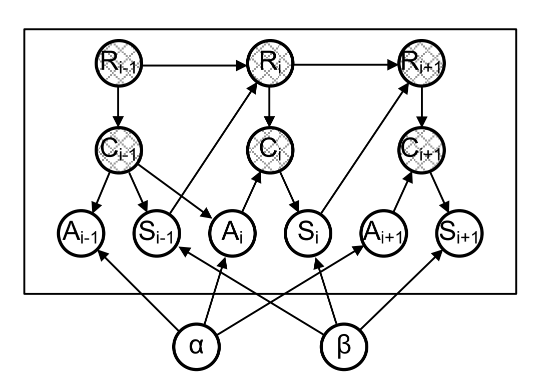


图1. 基于动态贝叶斯网络的查询效用性模型

3 大数据稀疏话题建模

针对网络大数据的分类、聚类、检索等基本的数据处理任务，都要求计算机能够自动挖掘出蕴含在数据中的语义信息。话题模型能够从繁杂的数据中提取与语义相关的低维的表达，从而被广泛应用于数据挖掘与处理的领域。传统的话题模型认为文档是由一组不同的话题产生，同一文档集共享相同的话题。然而在实际中，人们发现网络数据往往包含非常多的话题，而具体到某一篇文档，却仅仅覆盖了少量的一些话题。此外，很多社交网络、即时通讯中产生的大量数据中有一个共同的特点，即单个文本都非常的简短。这些问题导致传统的话题模型在计算性能和计算效率上都面临极大的挑战。本节中，我们分别从语义稀疏[9]、特征稀疏[10]角度展开研究，通过新颖的稀疏话题模型来解决大规模文档话题建模的挑战。

3.1 面向语义稀疏的成组稀疏话题模型

语义稀疏是指在大规模文档集中尽管存在大量的话题，但实际单个文档却往往覆盖极少的话题的现象。为解决这类话题稀疏的建模问题，同时也进一步提高话题计算效率和存储性能，近年来，很多学者都试图将稀疏性约束加入到话题模型的建模过程中，但效果却不甚理想。一方面，概率话题模型（诸如 PLSA[11] 、LDA[4]），对文本的建模方式清晰明确，每篇文档的话题表达可以被看成各个话题的组成比例，便于计算各篇文档之间相互关系。但是，传统的稀疏性方法（比如 LASSO [6]）却由于受到概率一致性约束的影响，而难以直接作用在

[6] Least absolute shrinkage and selection operator,最小绝对值压缩和选取,Lasso 的基本思想是在回归系数的绝对值之和小于一个常数的约束条件下，使残差平方和最小化

文档的话题成分上。另一方面，非概率话题模型（诸如 lsi[12]、 nmf[13] 、sparse coding[14]），不再要求文档表达具有一致的尺度，故稀疏性约束可以方便地控制文档的低维表达中 0 元素的个数。但这类方法破坏了话题模型的解释性，往往会造成度量数据之间关系的偏差，从而损害应用效果。

基于上述的认识，我们提出了一种结合概率话题模型的可解释性和非概率话题模型的稀疏性的新型话题模型——成组稀疏话题编码（Group Sparse Topic Coding）。该模型通过稀疏编码的思想，对文档中单词出现的次数进行建模。通过泊松分布和二项分布之间的关系，我们可以方便地从单词的话题编码推导出文本中各个话题所占的比例成分。另外，我们通过 group LASSO（成组 LASSO）的方式，将稀疏性约束直接施加于各个单词的编码之上，并将每篇文档中各个单词的编码的稀疏性对齐，从而达到控制文档话题稀疏性的目的。

给定一个文档集合 $\boldsymbol{D}$， 其中有 $\boldsymbol{M}$ 篇文档，$\boldsymbol{d}=\{w_1,w_2,\dots,w_n\}$，$\boldsymbol{\beta}\in$□$^{K+N}$，$\boldsymbol{K}$ 是话题个数，$\boldsymbol{N}$ 是词典大小，$s_{d,n}\in$ □^{K} 是文档 $\boldsymbol{d}$ 中单词 $\boldsymbol{n}$ 的编码。通过使用泊松分布对文档中单词出现次数建模，可以将单词的编码和单词的出现次数联系起来,具体的文档产生过程如下所示：

1. **For** $k\in\{1,2\dots,K\}$:对每个话题

   对一个词编码矢量 $s_k\in$□~M-Laplace(λ)采样

2. **For** $n\in I$: 对每个观察到的词

   For $k\in\{1,2\dots,K\}$:对每个话题

   对一个隐含的词频计数 w_{nk}~Poissan（s_{nk}, β_{kn}）采样

3. 得到词频计数 $w_n=\sum_{k=1}^{K} w_{nk}$

其中多元拉普拉斯（Multi-Laplacian）分布可以定义如下：

$$\text{M-Laplace}(s|0,\lambda^{-1})\propto \lambda N/2\exp(-\lambda\|s\|_2)$$

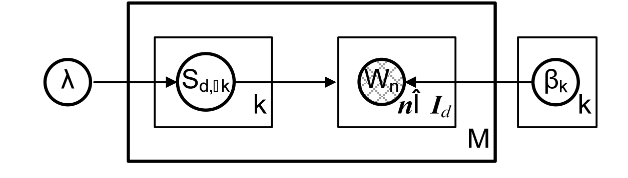


图2. 成组稀疏话题编码的概率图模型

用泊松分布产生单词次数的形式如下所示：

$$\text{Poisson}(w_{nk}|s_{nk}\beta_{kn})\propto(s_{nk}\beta_{kn})^{w_n}\exp(-s_{nk}\beta_{kn})$$

相应的概率图模型如图 2 所示。基于如上的文档产生过程，我们可以得到如下的优化问题

$$\begin{aligned}
&\min_{\Theta,\beta} L(\Theta,\beta)=-\ln P(\Theta,\beta|D)\\
&=\min_{\Theta,\beta}\sum_{d=1}^{M}\sum_{n=1}^{|Id|} l(s_{d,n},\beta)+\sum_{d=1}^{M}\sum_{n=1}^{K}\lambda\|s_{d,n}\|_2+C\\
&=\min_{\Theta,\beta}\sum_{d=1}^{M}\sum_{n=1}^{|Id|}\left\{\sum_{k=1}^{K} s_{d,nk}\beta_{kn}-w_{d,n}\ln(\sum_{k-1}^{K} s_{d,nk}\beta_{kn})\right\}+\sum_{d=1}^{M}\sum_{k=1}^{K}\lambda\|s_{d,n}\|_2+C
\end{aligned}$$

约束条件为：$S_{d,n}\geq 0,\forall d,n\in I_d \qquad \sum_{n=1}^{N}\beta_{kn}=1,\forall k$

可以看出在目标函数$\Lambda(\boldsymbol{\Theta},\boldsymbol{\beta})$中，多元拉普拉斯分布转化为 group LASSO，从而起到对每一篇文档中单词编码稀疏性的一致控制。针对该目标函数，我们采用坐标下降（coordinate-descent）算法迭代优化编码 $\boldsymbol{s}_{d,n}$ 以及话题分布 $\boldsymbol{\beta}$。基于学习得到的编码 $\boldsymbol{s}_{d,n}$ 以及话题分布 $\boldsymbol{\beta}$，我们可以恢复出文档中话题 $\boldsymbol{k}$ 所占的成分 $\boldsymbol{\theta}_k$k

我们通过在 20newsgroup 上的实验来验证所提方法的有效性。从实验结果我们发现，成组稀疏话题模型有着更强的稀疏控制能力，在文本分类精度上随着话题数目的增多，分类效果明显好于传统的话题模型 LDA 和 PLSA 以及最新的有监督话题模型 STC[15]。通过对比不同模型的训练时间，可以看出在计算速度上，我们的模型也显著优于传统概率话题模型。

3.2 面向特征稀疏的双词话题模型

短文本是互联网上一种常见的信息载体，如网页标题、文本广告、图像描述等。 近年来，随着社交网络的信息日益增多，如微博、状态信息、问答系统中的问题等短文本更是逐渐成为互联网上信息传播的主流媒介之一。传统的话题模型如 PLSA 和 LDA 面对特征稀疏的短文本数据会遇到以下问题：

- 短文本中大部分词都只出现一次，因此词频信息对词相关性和重要性判断不能产生区分度
- 由于短文本文档过短，上下文信息的缺乏会给一些二义性的词的话题判别带来困难

为了解决这些问题，我们提出了一种新的双词概率话题模型（Biterm Topic Model， BTM)来解决特征稀疏的文本话题建模难题。双词概率话题模型的出发点是直接从词共现关系去学习话题。这样做的好处是：（1）.词共现关系包含了上下文信息，比单个词更容易判断其中词的话题属性；（2）.词共现关系与文档长短无关。虽然单个文档内部的词共现关系比较稀疏，但只要数据足够多，全局的词共现关系仍然很充分。

给定一个短文本语料，我们首先从中抽取所有共现词对。在统计自然语言处理当中， 一个基本假设就是两个词共现的次数越多，它们的语义越相关。为了描述这种共现关系， 我们定义双词（***biterm***）为一个无序的词共现对。在短文本当中，由于文档长度短，文档中的主题比较集中，所以我们抽取其中任意两个不同的词组合构成一个双词。

双词概率话题模型为文档集合当中每个双词的产生过程建模。它假设整个文档集合是一个话题的混合分布，其中每个双词都来自于同一个话题，并且每个双词中的两个词关于该话题条件独立。双词概率话题模型的产生式过程描述如下（图模型如图 3 所示）：

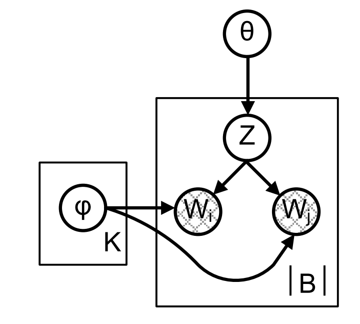


图3. Biterm 话题模型的概率图模型

1. 针对每个话题 z
   a) 基于狄利克雷分布 Dir($\boldsymbol{\beta}$)采样得到该话题下的词分布 $\boldsymbol{\varphi_z}$
2. 基于狄利克雷分布 **Dir($\boldsymbol{\alpha}$)**采样得到整个数据集的话题分布 θ
3. 针对双词集合 $\boldsymbol{B}$ 中的每个双词
   a) 基于多项式分布 **Multi($\boldsymbol{\theta}$)**对一个话题 z 采样
   b) 基于多项式分布 **Mulit($\boldsymbol{\varphi_z}$)**分别对两个单词 $\boldsymbol{w_i,w_j}$ 采样

根据以上产生式过程，一个双词 $\boldsymbol{b}=(\boldsymbol{w_i}, \boldsymbol{w_j})$的联合概率分布可以写成：

$$P(b)=\sum_z P(z)P(w_i|z)P(wj|z)=\sum_z \theta_z\phi_{i|z}\phi_{i|z}$$

整个语料的似然函数：

$$P(B)=\prod_{i,j}\sum_z \theta_z\phi_{i|z}\phi_{i|z}$$

双词概率话题模型可以通过吉布斯（Gibbs）采样算法求解其参数 $P(z)$和 $P(w|z)$。得到了这两个参数的值后，我们可以再利用以下方式估计文档的话题成分 $P(z|d)$：

$$P(z|d)=\sum_b P(z|b)P(b|d)$$

我们通过实验验证了我们提出的双词概率话题模型在短文本语料上的效果。实验基于公开的 TREC 2011 Microblog tracks 数据，包含了从 twitter（推特）采样两周的微博。对比方法包括 LDA，LDA-U[16](即将每个用户所有的微博汇总成一个文档后的 LDA)，Mix[17](Mixture of unigrams, 它假设每个文档只有一个话题)。为了评价学习到的主题质量，我们评价了话题的一致性得分（coherence score）C 如下式所示：

$$C(z,V^{(z)})=\sum_{t=2}^{T}\sum_{l=1}^{t}\log\frac{D(v_m^{(z)},v_l^{(z)})+1}{D(v_l^{(z)})}$$

其中，v 表示单词，$D(v)$表示词 v 出现过的文档数，$D(v,v')$表示词 v 和词 v'共现的文档数目。

从表 2 可以看出，双词概率话题模型（BTM）的一致性得分显著优于其他方法，说明双词概率话题模型学习到的话题更紧凑，可读性更好。

表2. Twitter 数据上的一致性得分结果

方法	5	10	20
LDA	-55.0±0.4	-236.4±2.0	-1015.7±5.9
LDA-U	-54.2±0.8	-234.8±1.1	-10009.4±4.4
Mix	-53.8±0.1	-233.0±1.4	-1007.6±6.7
BTM	**-52.4±0.1**	**-277.8±0.3**	**-990.2±3.8**

另一方面，为了评价文档话题表示的效果，我们从 twitter 数据中抽取了 50 个含有明确意义的哈希标签（hashtag），再将包含这些哈希标签的微博抽取出来，组成测试集。通过把每个哈希标签看成是一个类，我们计算了平均类内距离和平均类间距离的比值（H Score）。H 值越低，说明文档表达得越好。表 3 给出了各方法的 H 值，可以看出我们的方法双词概率话题模型显著优于其他方法。(其中*，**，***分别对应显著性检验 t-test 中的显著性水平 p-value 为 0.1,0.01 和 0.001）

表3. 不同话题模型的 H 值

方法	H Score	显著差
LDA	0.576 ±0.007	
LDA-U	0.564 ±0.011	>LDA*
Mix	0.503 ±0.008	>LDA-U**> LDA***
BTM	0.474 ±0.005	>Mix***>LDA-U**> LDA***

4 大数据下的排序学习技术

为了解决用户需求空间和网络数据空间的智能匹配问题，多种类型的排序模型得到广泛研究。其中排序学习技术通过机器学习的方法进行排序，是当前一类主流的排序模型。然而，传统的排序学习技术依赖于对全集样本的多级标注和学习，标注代价高且不能很好地体现检

索中关注位置的特点。如何提高排序学习技术在大数据下的性能成为了一个非常实际的课题。在本节，我们介绍 Top-k 排序学习框架[26]，通过建立关注于 Top-k 位置数据样本的标注、排序建模和评价体系，实现更高效的适用于大数据的排序学习体系。

4.1 Top-k 标注策略

鉴于在信息检索中用户主要关注前几个结果排序，而传统的标注数据并不能反映这样的需求，我们提出了 Top-k 标注策略。一方面，该标注策略采用相对标注方法，不同于传统的绝对标注[18,19]，让用户的标注更加简易和可靠；另一方面，该标注策略产生 Top-k 全序的数据样本，既符合排序的特点，又能够更好地用局部高质量的标注样本来提供全局学习，改进大数据排序学习的性能。

（1） 标注算法

为了实现 Top-k 标注，我们采用了基于小顶堆的 Top-k 标注算法，描述如下：1）随机选择 k 个元素，根据用户相对标注的结果构建小顶堆，堆顶元素为 t；2）从剩余元素中任选一个元素 r 与堆顶元素 t 比较，根据用户的标注结果，更新小顶堆，直至所有元素都被比较过，至此小顶堆中的元素即为前 k 个元素；3）根据用户的标注结果采用小顶堆排序对前 k 个元素排序。（见图 4）

Input：（1）D（一个词集合）；（2）k，排列项数
Begin
  随机选取集合 D 中的 k 项，记作 D_k，在其上构建一个最小顶堆（min-heap）H_k，并标注相关度
  For $d \in (D\text{-}D_k)$ **do**
    判断文档对（d，$D_k[1]$）中何者的相关度更高
    If d 相关度高于 $D_k[1]$，**Then**
      $D_k[1]= d$
      按照相关度更新 D_k 上的 H_k
    End If
  End For
  将 H_k 排序以得出按降序排列的前 k 项，记为 L_D
  将$(D\text{-}D_k)$加到 L_D
End
Output：L_D

图4. 基于堆排序的 Top-k 标注算法

（2） 标注复杂度分析

所谓标注复杂度即：对于任一个文档集合大小为 n 的查询，为了得到一个高质量的标注结果，需要用户做出判断的次数。绝对标注（包括 3 级标注，5 级标注等）的复杂度为 $O(n)$。根据上面提到的 3 个步骤：初始化 k 个元素的小顶堆的标注复杂度为 $O(k)$，通过对剩余 $n\text{-}k$ 个元素与该小顶堆的堆顶元素的比较与调整获得前 k 个元素的过程的标注复杂度为 $O((n\text{-}k)\log k)$，获得前 k 个元素的全序所需的标注复杂度为 $O(k\log k)$。因此基于小顶堆的 Top-k 标注策略的标注复杂度为 $O(n\log k)$。

（3） 数据集与标注流程

采用 TD2003（Topic Distillation task of TREC2003）中的所有查询，共 50 个，为了减小标注代价及实验方便，随机抽取了每个查询下的 50 个文档，保证至少一个是相关的。我们构造了一个可视化的标注工具，由 5 个人参与了标注，我们的标注方法确保每个查询都包含来自于两个不同的标注者的 Top-k 标注结果和 5 级标注结果。

（4） 实验结果

表4. 标注时间复杂度

方法	每次判断所用时间（秒）	每次查询所用时间（分）	判断复杂度	每次查询的判断数
Top-k 标注	5.51	13.13	$O(n\log k)$	142.76
五级标注	13.87	11.78	$O(n)$	50

表5. top-k 和五级标注的标注质量

(a) Top-k			
	$A \succ B$	A□B	$A \prec B$
$A \succ B$	0.6749	0.2766	0.0485
A□B	0.1138	0.8198	0.0664
$A \prec B$	0.1047	0.3779	0.5194

(b)五级标注			
	$A \succ B$	A□B	$A \prec B$
$A \succ B$	0.6272	0.2913	0.0815
A□B	0.2825	0.5232	0.1944
$A \prec B$	0.1534	0.3826	0.4640

由表 4 可知，五级标注与 Top-k 标注的时间复杂度相当。由表 5 可知，Top-k 标注的一致性要高于五级标注，即不易产生噪音。也就是说本文提出的基于相对标注的 Top-k 标注策略能够在保证时间复杂度基本不变的条件下，提高标注的质量。

4.2 Top-k 排序学习算法

如前所述，Top-k 序可以由不同的排序算法产生。不同的算法描述了 Top-k 序的不同的产生过程，这里我们对比了顺序产生和层次产生两种不同的方法。基于对 Top-k 序层次产生过程的认识，我们分别提出了产生式与区分式排序学习模型。

（1） 两种不同的产生 Top-k 序的方式

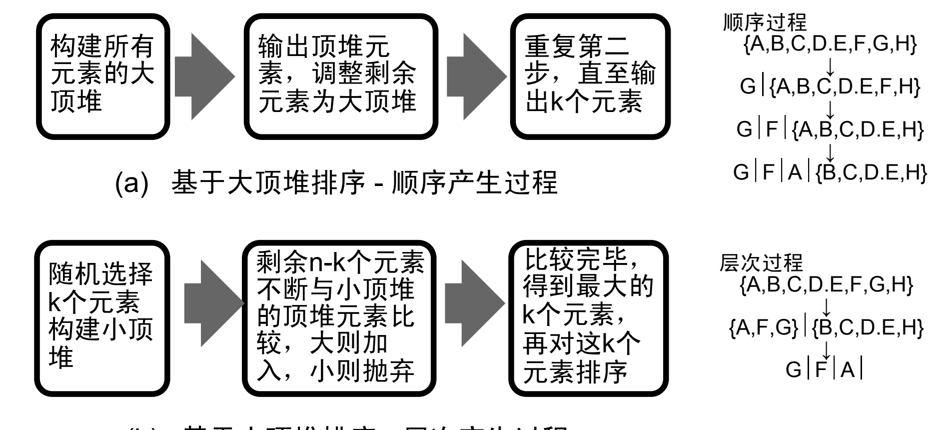


图5. 两种不同的产生 Top-k 序的方式

传统的 Top-k 排序学习模型侧重于从顺序产生过程来描述 Top-k 序，如 Top-k ListMLE[20] 与 Top-k CPS[21]，但是这样做是否是一个好的选择还是一个问题。因为现有经典的排序学习算法告诉我们 Top-k 序有不同的产生过程，小顶堆排序就是这样的一个例子，它描述了 Top-k

序的顺序产生过程，如图 5 所示。

（2） 层次产生过程的理论依据

由于顺序产生过程计算效率高，在排序学习模型中得到广泛采用，但是这种方法缺少理论支持。本文提出的基于层次产生过程的 Top-k 排序学习模型具有良好的理论依据——分槽独立性（Riffled Independence）[22]。分槽独立性是一类新的序结构上的独立性假设，它从元素的集合属性描述序的产生概率取决于同一集合内部元素之间的序以及集合与集合之间的序关系，如下式所示。

$$h(\pi)=m_{AB}(\tau_{AB}(\pi))\times f_A(\pi_A)\times g_B(\pi_B)$$

其中，m_{AB} 描述的是集合间的序关系，f_A 描述的集合 A 内部元素之间的序关系，g_B 描述集合 B 内部元素之间的序关系。分槽独立性为本文提供了一种很自然的 Top-k 序的分解方式。

（3） 概率 Top-k 排序学习模型（HOM）

Top-k 序σ中的元素可以看作来自 T 和 F 两个集合，其中 T 集合中的元素都是和查询相关的，F 集合中的可以认为是与查询不相关的。因此 Top-k 序的产生概率可以分解为三部分的乘积。然而由于我们仅仅关注于前 k 个位置的序，对于 F 中的元素只知道是位于这 k 个元素后面的，而其中元素之间的序未知，我们也不关心，因此可理解为随机序。基于层次产生的思路，Top-k 序可以分解为两部分：（1）第一层为前 k 个元素与后面元素之间的序关系；（2）第二层 为前 k 个元素之间的全序关系。依据分槽独立性，Top-k 序σ的产生概率可用如下公式描述。

$$P(\sigma)=P_{TF}(T\prec F)\times P_{T}\times(\sigma_{T})$$

考虑到计算效率，对于第二层的前 k 个位置的全序关系σ_T 的产生概率采用传统的顺序产生模型描述，如鲁斯（Luce） 模型[23]等。对于第二层的集合间的序关系的产生概率，可以有多种描述方法，我们列举了其中的三种，分别记为 Group-to-Group（组对组），One-to-One（一对一），以及 One-to-Group（一对组）。组对组方式将集合 T 与 F 都分别视为一个整体。一对一方式从微观入手，进行两个集合中的元素之间的比较。一对组方式，将重要的一方 T 分解为元素，不重要的一方 F 整体对待。因此概率 Top-k 排序学习模型 HOM 有三种实现方式：1）组对组+Luce 记为 HOM-GG；2）一对一+Luce 记为 HOM-OO；3）一对组记为 HOM-OG。

（4） 区分式 Top-k 排序学习模型-聚焦排序（FocusedRank）

从似然函数的角度来看，HOM 的似然损失函数可以看作是两部分损失之和，一部分是前 k 个元素之间的全序关系的成表（listwise）损失，一部分是前 k 个元素与后面元素两两之间的配对损失函数。因此可以进一步推广为 T 上成表损失与 T 与 F 之间的配对损失的线性组合。这里将 T 与 F 之间构建的所有“配对”的集合记为 P。因此，区分式 Top-k 排序学习模型的损失函数如下式所示

$$L(f,q_i)=\beta\times L_{\mathrm{list}}(f;T_i,y_i)+(1-\beta)\times L_{\mathrm{pair}}(f;P_i,y_i)$$

其中 L_{list} 和 L_{pair} 分别表示成表损失函数和成对损失函数，β 表示权重系数，y_i 表示标注信息。

这样排序学习中研究得较多的配对算法与列表（listwise）算法都可以应用到区分式排序中，但是考虑到物理意义，本文提出了以下三种组合方式：（1）基于支持向量机的聚焦排

序（FocusedSVM：SVM-MAP+RankSVM）；（2）基于神经网络的聚焦排序（FoceseNet: ListNet+RankNet）；（3）基于集成学习的聚焦排序（FocusedBoost：AdaRank+RankBoost）。这样区分式排序可以采用不同的优化算法求解：基于支持向量机的聚焦排序采用支持向量机（SVM）中常用的割平面的优化算法来求解，基于神经网络的聚焦排序采用梯度下降求解，基于集成学习的聚焦排序则可采用 Boosting[7]的过程求解。

4.3 Top-k 序的评价指标

针对绝对标注的评价指标 NDCG[24]与 ERR[25]得到广泛的应用。本文针对 Top-k 序标注对其进行扩展，主要解决方法是将位置信息映射为相关性程度。Top-k 序σ定义在 $D=\{x\}_{i=1}^{n}$上，假设位置 k 之后的元素的位置均为 k+1，则元素 x_i 相关程度可以简单定义为 $y_i=k+1-\sigma(x_i)$。因此本文提出了针对 Top-k 标注的 NDCG 与 ERR 记为 K-NDCG 与 K-ERR

$$K\text{-NDCG@l}=\frac{1}{N_l^{best}}\sum_{i=1}^{l}\frac{2^{y_i-1}}{\log_2(1+i)}$$

$$K\text{-ERR}=\sum_{i=1}^{n}\frac{1}{i}R(y_i)\prod_{j=1}^{i-1}(1-R(y_i))=\frac{2^r-1}{2^{y_{\max}}}$$

（1） HOM 模型的性能

HOM 模型的提出是为了尽可能好地对具有 Top-k 序标注的训练数据的特征建模，因此实验中采用了 Top-10 MQ2007 与 Top-10 MQ2008 两个数据集，分别是 MQ2007-list 与 MQ2008-list 的子集，取得 Top-10 标注来模拟具有 Top-k 标注的数据集。实验结果如下所示。实验中对比了现有的主流概率 Top-k 排序学习算法 Top-k ListMLE 与 Top-k CPS。实验结果如图 6 所示，图中纵坐标为排序中考虑位置因素的相关性指标。从图中可以看出 HOM 模型显著好于两个基准方法。

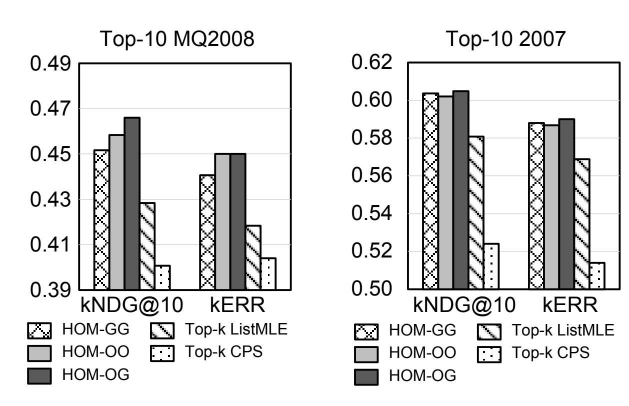


图6. HOM 性能对比

（2） FocusedRank 的性能

采用的实验数据集有两个：(1)MQ2007-based；(2)TD2003-based。对于这两个数据集合分别考察了 FocusedRank 在三级标注上 Graded MQ2007（3 级），Graded TD2003（5 级）与 Top-10 标注上 Top-10 MQ2007，Top-10 TD2003。其中 Graded TD2003 与 Top-10 TD2003 是采用前面提到的标注工具实际标注得到的结果；Graded MQ2007 是采用现有的 LETOR 中的数据，Top-10 MQ2007 是根据 MQ2007-list 进行人工合成后得到的结果。实验结果如下表 6 与表 7 所示，可以很明显看到，本文提出的 FocusedRank 算法在具有 Top-k 标注的数据集上取得较优的性能，在多级标注数据集上与传统的 Listwise 算法与 pairwise 算法也是可比的。

[7] 一个将弱学习(weak learn)算法融合为强学习算法(strong)的方法,基本思想是将多个能力较弱的分类器迭加得到一个更强的分类器

表6. FocusedRank 在 MQ2007 上的性能对比

方法	Graded MQ2007		Top-10 MQ2007	
	NDCG@10	ERR	***K***-NDCG@10	***K***-ERR
SVM^{MAP}	0.4419	0.3146	0.6690	0.6227
RankSVM	0.4447	0.3178	0.6655	0.6205
FocusedSVM	0.4400	0.3196	**0.6739**	0.6287
AdaRank	0.4345	0.3061	0.6190	0.5637
RankBoost	0.4438	0.3101	0.6571	0.6131
FocusedBoost	0.4422	0.3199	0.6628	0.6187
ListNet	0.4442	0.3206	0.6613	0.6195
RankNet	0.4451	0.3157	0.6603	0.6143
FocusedNet	**0.4459**	**0.3223**	0.6735	**0.6336**
Top-k ListMLE	0.4443	0.3168	0.6673	0.6228

表7. FocusedRank 在 TD003 上的性能对比

方法	Graded TD2003		Top-10 TD2003	
	NDCG@10	ERR	***K***-NDCG@10	***K***-ERR
SVM^{MAP}	0.5801	0.5102	0.3858	0.3829
RankSVM	0.5991	0.5072	0.3872	0.4025
FocusedSVM	0.5885	0.5129	0.3886	0.4041
AdaRank	0.5982	0.5132	0.3766	0.3777
RankBoost	0.5750	0.5119	0.3789	0.3970
FocusedBoost	0.5955	0.5466	0.3980	0.4214
ListNet	0.5985	0.5225	0.3624	0.3962
RankNet	0.5983	0.5059	0.3813	0.4199
FocusedNet	**0.6082**	**0.5660**	**0.4058**	**0.4603**
Top-k ListMLE	0.5885	0.5083	0.4007	0.4028

5 结束语

为了帮助人们从网络大数据中有效发现和准确定位所需的信息，信息检索与挖掘技术面临着全新的挑战。本文介绍了我们在面向网络大数据的深度检索与挖掘技术方面展开的系列研究以及取得的相关成果，包括：基于大规模用户查询日志的查询理解，大数据稀疏话题建模，以及大数据下的排序学习技术。未来，我们将继续从这几个方面展开深入研究，利用大数据环境进一步改进对用户查询、文档数据的理解，提高排序在大数据下的性能。

参考文献：

[1] Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li, Named entity recognition in query, Proceedings of *the 32nd international ACM SIGIR conference*, 2009, 267-274.

[2] Jiafeng Guo, Xueqi Cheng, Gu Xu, Xiaofei Zhu, Intent-Aware Query Similarity, Proceedings of *The 20th ACM Conference on Information and Knowledge Management*, 2011, Glasgow, UK, Oct. 2011.

[3] Xiaofei Zhu, Jiafeng Guo, Yanyan Lan, Xueqi Cheng, More Than Relevance: High Utility Query Recommendation By Mining Users' Search Behaviors, Proceedings of *The 21th ACM Conference on Information and Knowledge Management*, 2012.

[4] David M. Blei , Andrew Y. Ng , Michael I. Jordan, Latent dirichlet allocation, *Journal of Machine Learning Research*, 2003.

[5] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In Proceedings of *the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '00*, pages 407–416, New York, NY, USA, 2000. ACM.

[6] N. Craswell and M. Szummer. Random walks on the click graph. In Proceedings of *the 30th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '07*, pages 239–246, New York, NY, USA, 2007. ACM.

[7] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna. The query-flow graph: model and applications. In Proceedings of *The171th ACM Conference on Information and Knowledge Management*, pages 609–618, 2008.

[8] Xiaofei Zhu, Jiafeng Guo, Xueqi Cheng, Pan Du, Hua-wei Shen. A Unified Framework for Recommending Diverse and Relevant Queries, Proceedings of *the 20th ACM International World Wide Web Conference*, Hyderabad, India.

[9] Lu Bai, Jiafeng Guo, Xueqi Cheng, Yanyan Lan and Wolfgang Nejdl. Group Sparse Topical Coding: From Code to Topic, In Proceedings of *the Sixth ACM WSDM Conference*, Rome, Italy, 2013.

[10] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xuiqi Cheng. A Biterm Topic Model for Short Texts. In Proceedings of *the 22nd international conference on World Wide Web*, Rio de Janeiro, Brazil, 2013, ACM.

[11] T. HOFMANN. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning Journal*, 2001.

[12] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W.Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of The American Society for Information Science and Technology*, 41(6):391–407, 1990.

[13] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, Oct. 1999.

[14] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, June 1996

[15] J. Zhu and E. P. Xing. Sparse topical coding. *Uncertainty in Artificial Intelligence*, pages 831–838, 2011.

[16] L. Hong and B. Davison. Empirical study of topic modeling in twitter. In Proceedings of *the First Workshop on Social Media Analytics*, pages 80-88.ACM, 2010

[17] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2):103-134, 2000.

[18] Robert Burgin. Variations in Relevance Judgments and the Evaluation of Retrieval Performance, *IPM*, 28(5), pages 619-627, 1992.

[19] Jaana Kekalainen. Binary and Graded Relevance in IR Evaluations-Comparison of the Effects on Ranking of IR Systems, *IPM*, 41(5), pages 1019-1033, 2005.

[20] Fen Xia, Tie-Yan Liu and Hang Li. Statistical Consistency of Top-k Ranking, In Advances in *Neural Information Processing Systems* 22 (2009), pages 2098-2106.

[21] Tao Qin, Xiubo Geng and Tie-Yan Liu. A New Probabilistic Model for Rank Aggregation, In *Advances in Neural Information Processing Systems* 22 (2010), pages 1948-1956.

[22] Jonathan Huang, Carlos Guestrin. Riffled Independence for Ranked Data, In Advances in *Neural Information Processing Systems* 22 (2009).

[23] J. I. Marden. Analyzing and Modeling Rank Data. *Chapman & Hall*, 1995.

[24] Kalervo Jarvelin, Jaana Kekalainen. Cumulated gain-based evaluation of IR techniques, *ACM Transactions on Information Systems* 20(4), pages 422–446, 2002.

[25] Olivier Chapelle, Donald Metlzer, Ya Zhang and Pierre Grinspan. Expected Reciprocal Rank for Graded Relevance. Proceedings of *the 18th ACM Conference on Information and Knowledge Management,* pages 621-630, 2009.

[26] Shuzi Niu, Jiafeng Guo, Yanyan Lan and Xueqi Cheng. Top-k Learning to Rank: Labeling, Ranking and Evaluation, Proceedings of *the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 751--760, 2012.

作者简介：

郭嘉丰： 中国科学院计算技术研究所、网络数据科学与工程研究中心基础研究部负责人、副研究员，guojiafeng@ict.ac.cn

程学旗： 中国科学院计算技术研究所、网络数据科学与工程研究中心主任、研究员